DO RATE AND VOLUME MATTER? TRANSACTION COST LIMITS TO ECONOMIES OF SCALE

Cotton M. Lindsay & Michael T. Maloney
Department of Economics
Clemson University

In the traditional treatment, economies of scale are attributed to a hodgepodge of sources. A typical list might include Adam Smith's famous "division of labour," economies of large machines, the integration of processes, massed reserves, and standardization. The list can be partly systematized because, when considered in detail, these various economies are themselves the result of various other more basic and occasionally overlapping principles. For example, both economies of massed reserves and economies of standardization are to a certain extent the product of the statistical “law of large numbers.”

However, even this analysis fails to strike to the heart of the matter because the technological factors however described that reduce costs with scale do not in themselves imply that large firms can produce at lower cost than small firms. The possible presence of such technological scale economies does not give us adequate knowledge to predict the structure of industry. These forces of nature may combine to make it cheaper to get things done in big chunks. However, this potential will be economically important only in the presence of transactions costs.

Firms can specialize their production processes and hire out the jobs that require large scale. Realistically all firms hire out some portion of the production process regardless of their size. General Motors ships many of its automobiles by rail, but does not own a railroad for this purpose. Anaconda uses a great deal of fuel oil in its production of copper, but it does not own oil wells or refineries. By contracting for the services of facilities that they cannot fully exploit alone, even small firms in an industry may thus enjoy the full measure of scale effects for *all* processes it employs. In principle at least, it is possible for average costs to be equally low for large firms as well as small firms that cannot use large scale technology to capacity. The forces that determine this rent-or-buy choice are the important conditions for assessing the practical effect of economies of scale.

The purpose of this paper is to explore the way transactions costs affect economies of scale. The paper is in part an exercise in history of thought because it follows the development of the notion of economies of scale through the work of many great scholars who have struggled with the problem. This paper is separated into four parts. The first two analyze the traditional arguments about increasing returns to scale. The third uses the taxonomy of rate and volume to catalog economies and diseconomies. Finally, the fourth section builds a model of the transactions costs that prevent firms from fully exploiting economies of scale.

I. Smith & the Division of Labor

The seminal discussion of the notion of returns of scale comes from Adam Smith in his classic example of the pin factory. As he tells the story, the average, uninitiated man could, at his best, produce twenty pins a day. However, in the business, it was routine for a small shop to produce 6000 or more pins per worker, per day. This gain was (and is) obtained by worker specialization or the division of labor.[1] Smith writes

> One man draws out the wire, another straights it, a third cuts it, a fourth points it, a fifth grinds it at the top for receiving the head; to make the head requires two or three distinct operations; to put it on is a peculiar business; to whiten the pins is another; it is even a trade by itself to put them into the paper; and the important business of making a pin is, in this manner, divided into about eighteen distinct operations, which in some manufactories, are all performed by distinct hands.

Smith enumerates three “circumstances” through which specialization operates to produce these efficiency gains. First, dexterity improves as processes are repeated. Second, time is saved by reducing the amount movement from one work station to the next. Finally, many workers increase the likelihood that new devices and ways to save labor effort will be invented. This last phenomenon may have been more common in Smith's day than it is in the space age we now inhabit.[2]

The Dexterity of Labor

The dexterity to which Smith referred can be interpreted as the memorization of a set of motions that allows the worker to perform them automatically. Often workers perform best when they enter an almost trance-like state. Consider typing. Good typists do not think about the individual letters that compose a word; they type words. Their fingers unconsciously perform this task without conscious direction. It is possible that effects of repetition are achieved by a process similar to theorizing. Repetition produces speed by allowing the mind to adopt a model of the tasks to be performed just like a scientist builds a model of a physical phenomenon he seeks to explain. Repetition refines the mind's model until it becomes almost a reflex. Only particular signals trigger the correct sequence of motions.[3] The worker is free of the involvement, doubts, and creative impulses that are associated with conscious willing of each task. An economy of scale results because people can master only a limited number of such tasks.[4] Thus, production must be allocated among several individuals and the minimum efficient plant size is fixed by the output that will fully occupy these workers acting together.

When dexterity spawned by repetition occurs with teams of workers, it is called the learning curve phenomenon. All inputs develop interactive efficiencies by repetition of the process. In sailing, as an example, the coming-about drill is relatively simple; the learning curve flattens out rapidly. On the other hand, dropping the spinnaker rounding a mark takes substantially more practice to develop the coordination of the crew. The learning curve is flat when all the inputs necessary for production are used in unconscious harmony. This comes from practicing an exercise as a team until each team members can do their job in their sleep.

[1] Smith, p. 4.

[2] It would be interesting to know how many innovations in recent times have resulted from assembly line workers suggestions. One of our Croatian graduate students informs us that such innovations were reported in the press in communist Yugoslavia somewhat frequently.

[3] How often does it happen in these kind of habitual processes that if the activity is interrupted in the middle you have to start over completely?

[4] The telephone company has done research on how much the average person can memorize. In piece rate sewing, the number of tasks may be very large for some items, and while assembly line construction is not typically feasible, the best seamstresses still achieve the meditative state.

Even where workers can in principle master all tasks on the production line, labor costs can typically be lowered *per task* by employing workers who have mastered only a few. Workers with many skills must be compensated for acquiring them all; they must be paid more than workers with only a few. However, they can use these skills only one at a time. Another way of putting this is that unspecialized labor is wasteful of human capital. When the welder puts down his torch and picks up a paint brush, his skills in the former task are wasted. He cannot do both at the same time, so there is no reason to require him to do both *unless the amount of work is insufficient to keep him fully occupied*. If the volume of work is large enough to require a full-time painter and a full-time welder, however, people may be employed at lower wages who can only weld or paint. Thus even if the amount of labor input per unit of output is unaffected by labor specialization, labor costs per unit of output fall because the labor employed is cheaper. These economies of skill may be exploited increasingly as scale expands. As scale gets larger, workers may be fully occupied doing an increasingly more limited set of tasks. More highly skilled workers (i.e. those capable of doing a wider range of tasks) can be replaced by lower skilled (and lower paid) workers, and costs fall.

The perceptive reader will note that this is really a transaction cost argument. All the gains from specialization could be exploited by firms with a low rate of output, if they could hire specialized workers and productive facilities for sufficiently short periods of time. That is, to achieve the same degree of specialization at half the rate of output, the same number of employees might be hired for half the time. The reason this is not done is that it is costly for workers to divide their work time among more than one employer. We elaborate on these transaction cost arguments in section III below.

Assembly Lines.

Assembly lines themselves are sources of efficiency gains apart from making possible specialization of labor among workers. They make possible a savings achieved by avoiding moving workers from point to point. One aspect of this is the fact that the cost of the energy and coordination required to move assemblies along a track or conveyor is less than the cost of having the men move from one unit to the next in concert. The energy savings result from the economy of large machines (in this case large motors), a topic that we will take up shortly. Note, though, that moving the production pieces along an assembly line allows the supplies and tools to remain stationary, producing an additional economy. For example, if a computer terminal is composed of 1500 pieces that require 30 different tools to install, it may be much cheaper to spread the supplies and tools out across 10 work stations, which average 3 tools and 150 units of material, rather than pile them all up at one spot. Costs are lower because of reduced confusion of parts and tools.

It is interesting to note that even in a robotic factory, assembly lines are still used. Robots are normally very specialized, welding one or two points, tightening three bolts, or cutting along a single line. The piece is passed from one robot to the next because it is cheaper to build many specialized robots and locate them at several stations than to build the "everyman" variety. The reasons for this are parallel to the arguments we have just discussed concerning the economies due to specialization of labor. Although robots, unlike people, can do several things at once, they are prevented by the constraints of physical space from performing too many tasks on the same unit at the same time. It therefore does not pay to design into robots the capacity to do more tasks than can be simultaneously performed. To do so will leave some of that capacity less than fully employed. Robots are subject to economies of skill, too. It is also cheaper to spread out the work stations and move the work to the robots, than to have robots moving from one work station to another.

Cross Currents

In spite of the insight Smith brought to the discussion, the phenomena that he analyzed do not operate monotonically to lower unit cost. No doubt in many settings, the firm can lower its cost by hiring a single-task worker. Even so, narrowly specialized labor can create some problems that tend to offset the cost savings. There is a broad movement today for firms to stress teams and team work. Basically, firms engage in training individuals to perform more than one task, and while working as a team, employees rotate their duties. They work on different tasks over a period of time. Rotation is said to help workers "feel more important," to increase their interest in work, and to break down the monotony of doing the same job over and over, which often leads to lack of motivation and low productivity. Thus, gains from specialization in the work force are increasingly being questioned as they tradeoff against the costs of apathetic employees.

II. Robinson & the Aggregation of Capital

There are other sources of economies of scale that are not related to the division of labor. Several of these are discussed by Robinson whose work contains the litany of commonly cited sources of these effects. The three that we discuss here are economies of massed reserves, economies of management, and the economies of large machines.

Massed reserves are akin to an insurance pool. Inventories of production devices, tools, pumps, fork lifts, and the like, can be minimized per unit of output without loss of backup potential where the facility is large. A numerical example may make this clearer. Assume that there are sixteen firms that make hand calculators. On average each firm receives orders for 8,000 per week and produces the same number. The main component of these devices is a microprocessor or chip. Due to assembly errors and the fact that many of these chips are defective, however, the number required for production of 8,000 calculators will be greater than 8,000 and will vary for each firm from week to week. Let us assume that on average, a firm will use 10,000 per week to produce 8,000 calculators. If each firm is to produce this number 95 percent of the time, it must maintain an inventory of chips, over and above the 10,000 it normally requires, out of which to meet this production target.

To know how large the inventory must be, we must know certain characteristics of the distribution of weekly chip requirements. Assume that the distribution is normal, and that the standard deviation in requirements to any single firm is 2432. With this information we can determine that the additional inventory necessary to produce the desired number of calculators 95 percent of the time. This number is 4,000 chips. The 16 firms together must therefore hold a combined inventory of 64,000 chips.

Now assume that the 16 firms merge and fill all their chip requirements from a common inventory. When some firms have an unusually high number of defects, others will have fewer than average, so the common inventory will be less than the 64,000 required for 16 individual firms. The precise size of the required inventory in this case may be determined by finding the number that will assure that the *average* needs of the combined 16 firms are no more than 14,000 chips 95 percent of the time. Our assumptions about this underlying distribution allow us to determine that the combined inventory must contain only 16,000 of these components.[5]

Another, less convincing scale effect discussed by Robinson is economies of management. He claims that management is affected by the division of labor principle just like workers. We find this a little hard to believe at least within the context of our interpretation of Smith's idea. Management is the resource that monitors production, which is a task not easily performed in a trance. Robinson further claims that some managerial functions are similar to start-up costs. He uses the example of a sales forecasting staff in which as output doubles, the size of the unit increases but does not double. Whether this is the common state of nature or not is hard to say. Casual observation indicates that the larger a firm becomes, the more different divisions of management it has, like advertising, marketing, sales forecasting, corporate planning, etc. Thus, even if any one grows more slowly than output, increasing the number may defeat any managerial economies.[6]

Economies of management may have some validity in the organization of production. The management of a large firm can schedule activities so that work proceeds with a minimum of interruption. Tools can be fetched, bits can be sharpened, supplies can be tended, debris can be removed, all so that production continues unaltered. This can happen in a big firm or small, but one suspects that specialized crews of attendants require some minimum size to keep them busy. This effect is like two others that Robinson coins phrases for, the integration and balance of processes. Both of these are analyzed at length by Stigler. Consider each component function of a production process as having its own average cost function. The firm's average cost function is simply the amalgamation of these. Some processes will require a substantial output level to achieve minimum average cost, while others may be in the increasing cost range by this time. Altogether, integrating and balancing these forces will be arguably cheaper for the large firm than for the small.

All things considered, the effects of management are usually held to be diseconomy effects rather than the opposite. As the size of the firm becomes large, things begin to fall through the cracks. The ability of upper management to juggle the goings-on of the many divisions declines as the firm size increases. Hence, unit costs rise. The problem consists of setting up the proper incentive, reward, and monitoring system for the managers of the many divisions, plants, and assembly lines. If the large firm were to spin off into many smaller firms, the market would sanction the behavior of the managers directly. This scrutiny is what must be replaced when the pieces are assembled under the aegis of central control.

The last element of the economies of scale taxonomy developed by Robinson is economies due to large machines. These are widely referred to elsewhere as engineering economies and refer to such properties as the strength of metal parts increasing more rapidly than their thicknesses and the carrying capacities of pipes increasing more rapidly

[5] Ninety-five percent of the time firms will require less than $10{,}000 + 1.645s_i$ where s_i is the standard deviation of the distribution. If the weekly usage is 10,000, the inventory required for each of the 16 firms to satisfy all of its needs 95 percent of the time is therefore given by the term $1.645s_i = 4{,}000$. Where inventories are combined, all requirements will be filled so long as the *average* usage does not exceed $10{,}000 + 1.625s_i/2!n$ where $n = 16$. The value of this second term, the required average inventory, is 1,000 chips. The total required inventory for the combined firm is therefore 16,000.

[6] A word of caution is in order. The economies of scale discussion is a ceteris paribus experiment. We examine the change in average cost as output change, holding everything else constant. Hence, it is not really accurate to look across industries where the type of output changes. Management is the resource that monitors the application of other resources. The difficulty of performing this monitoring function will change across industries.

than the materials required for their fabrication. There are principles of the physical world that operate to produce output increases with less than proportional increases in materials. Containers are exemplar. The volume of a container increases with cube of the dimensions while the material for its walls increases by only the square of those dimensions. For example, if the height and diameter are doubled, the capacity increases eightfold. However, the material of the walls will consume only a fourfold increase.

Ships and airplanes experience similar gains in efficiency with size. The energy required to move a boat through the water is proportional to the wetted surface of its hull. Sufficient force is required to overcome the friction of the water against the hull's surface, and that force is directly proportional to the area of the surface. This area increases at only the square of the interior dimensions, however, while the ship's carrying capacity expands with the cube of those dimensions. It is widely understood among sailors that large boats sail faster than small ones holding the shape constant.[7]

The cost reducing potential of standardization is a special case of this principle. Because of set-up and measurement costs it requires less time to cut out forty shirt sleeves at once than to cut out each one individually. Doing so makes necessary that each sleeve attach to a shirt from a matching pattern. Standardizing makes it possible to exploit the gains from large machines. The fundamental principle at work is best summarized by Alchian:[8]

> The method of production is a function of the volume of output, especially when output is produced from basic dies—and there are few, if any, methods of production that do not involve "dies." Why increased expenditures on more durable dies should result in more than proportional increase of output potential is a question that cannot be answered, except to say that the physical principles of the world are not all linear[.]

In addition, like economies of massed reserves, standardization permits firms to adjust to the uncertainties of the market at lower cost. As an example, General Motors has gone to the extreme of standardizing the frames of its cars. GM now has assembly plants that can handle any of three sizes so that production can be switched without stopping the line.

III. Alchian & Volume versus Rate

Alchian's discussion of the "physical principles of the world" is a jumping off place to generalizing the notion of economies of scale developed since Smith. Alchian summaries the state of the classical understanding about cost. In the form of five propositions, he creates a taxonomy of cost effects in terms of the relationship between rate and volume. Alchian's propositions are that costs increases at an increasing rate in terms of rate holding volume constant and increases at a decreasing rate with respect to volume holding rate constant. Put simply, volume effects create economies of scale and rate effects create diseconomies of scale.

The sense in which Alchian says volume creates economies of scale hinges on his "dies" argument. To plan to produce more, allows the firm to choose technologies that use more durable dies, hence yield lower average cost. But this same volume effect can encompass the division of labor. It is not logical for a worker to memorize a set of skills

[7] The exception is planing where the boat rises up and skims the surface of the water.

[8] Alchian, p.282.

that will be employed only once, or to improvise a jig for a one-of-a-kind piece of furniture.

In explaining that rate effects create diseconomies, Alchian is more arcane. In the limit, the classical model says rate generates increasing per unit cost because of specialized resources. That is, as rice becomes the only grain product in the world, the cost of flooding mountain tops becomes prohibitive. This explanation is not particularly appealing, however, when we attempt to explain why rice farms in southern Louisiana are typically 220 acres instead of 110 or 440.

The standard explanations, left unstated by Alchian, typically reference diseconomies of management and the crowding that occurs as workers try to produce a lot of output rapidly. Another of the physical principles of the universe is that machines run less efficiently at excessive speeds. The firm chooses the machine (die) that will wear out precisely at the end of the planned volume of the project. Volume determines the durability of the machine. Rate determines the speed at which it is run. The faster the machine runs, the more stress is created *per unit of output*. That is, holding volume constant, speed increases costs at an increasing rate. As a simple example, a car with a bad engine bearing will lock up if driven at 60 miles an hour for one mile, but the same car can be driven 60 miles or more at 10 miles per hour. Speed increases the severity of break-downs when they occur.

The "U-Shaped" Cost Curve.

Putting the rate and volume effects together gives us the classical U-shaped average cost curve. The most enlightening way to understand this implied relation is by describing it graphically. Figure 1 reproduces the three dimensional cost relationship that Alchian uses to describe his propositions on costs. As shown in Figure 1, looking across the volume (V) dimension, costs increase at a decreasing rate. On the other hand, holding volume constant while increasing rate (R) can be seen to increase costs at an increasing rate.

The marginal relations between cost, rate, and volume (Alchian's propositions) are embedded in Figure 1, but so too is an implied relation between rate, volume, and time. Rate and volume can be mapped to each other by time (t), i.e., $V = R \bullet t$. For any given value of time, choosing rate identically chooses volume. That is, if the time period of production is constant, increases in total output come from simultaneous increases in rate and volume. Holding time constant, increases in rate are implicitly increases in volume. Graphically this means that projecting a time ray along the floor of Figure 1 shows the rate-volume combinations that yield more output holding the production time period constant.

It is interesting to look at a slice of Alchian's three dimensional cost curve along a constant period-of-production time ray. This is shown in Figure 2. Let $t = t_0$ and, for sake of exposition, we show the three-dimensional shape truncated for values of R greater than V. This reveals the *ridge* that is the pattern that costs follow as rate and volume are simultaneously increased holding time constant at t_0. Note that the contour of this ridge has the classic cubic shape that yields U-shaped average cost. As production is expanded along this ridge, costs first increase at a decreasing rate and then at an increasing rate. This implies that average cost falls and then rises.

Alchian's five propositions imply that in the classical world where rate and volume are assumed to vary proportionately, the volume effect dominates first and then the rate effect takes over.[9] Of course, it may be that for a given project, the volume effect is so weak relative to the rate effect as to be empirically meaningless. It may also be the case that the volume effect is so strong that average cost continues to decline well beyond feasible economic levels of production.[10] What Alchian has done in a masterful fashion, is dissect economies and diseconomies of scale into volume and rate categories.

The Period of Production.

A note on the time dimension is in order. In Alchian's model, explicit decisions are made about scale in both dimensions. The prototype for this kind of production is building construction. Both the volume (size and number of structures) as well as the rate are negotiated in contracting process. The latter is determined by virtue of construction time or deadline being a part of the terms. That is, once the contract fixes volume and time, rate is uniquely determined. Much production seems to fall into this category. The planned runs of particular automobile models, clothing of a given design, television sets and computers are limited by considerations of eventual changes in technology and fashion.

On the other hand, the standard neoclassical firm has an implicit planning horizon of infinity. The planned volume of a such a firm is indefinitely large as some positive rate of output extends into the distant future. In this slightly implausible[11] but vastly useful simplification, we can imagine that the firm proceeds as follows. It scans over all possible rate/volume ratios to find the combination that produces the *lowest* minimum average cost. Then it adopts this rate/volume combination for the production plan. Every firm in the neoclassical industry produces at this rate year round.

Note, however, that the time dimension in which the rate of output is measured plays no real role in the determination of cost once the firm adopts the optimal rate/volume combination. The shape of the average cost curve for an arbitrary time period *t* is a compressed transformation of the shape for any larger multiple of *t*. The unit of time in the output dimension is arbitrary. For instance, if GM can produce 100,000 cars a month at a total cost of $1 billion, they can produce 50,000 cars in two weeks for $500 million. Unit costs are $10,000 in both cases. The volume effects, which results from durability and the division of labor, are still the driving force behind the economies of scale portion of any time-constant, rate-volume expansion, and the rate effects still dictate the onset of diseconomies.[12]

While this reconciliation of the Alchian and neoclassical theories of cost works well enough so long as the cost minimizing output leaves scope for many firms, it retains a handicap for the analysis of industry structure. It can be used only for the analysis of firms that produce at a constant rate per period. Yet, as we have just seen, the length of period

[9] It is curious that Alchian fails to take this final step. He refuses to claim that the combined effect of rate and volume can be predicted from the analysis.

[10] This seems to be the case in shipping. For reasons already discussed, larger ships always have lower costs. Only port space constrains their use. This raises the interesting paradox of why all transoceanic freight is not all shipped on one “titanic” vessel. At some point, the risk aversion associated with having all our transoceanic product on one ship may offset the classical rate diseconomy, but we seem to be far from that margin now.

[11] It is difficult to imagine any production process totally unconstrained by some implicit volume constraint. The gain from employing dies that yield 200 year's worth of the world's consumption of just about anything seems destined to be outweighed by the expected obsolescence of the present production technology.

[12] Thus, each average cost function defined for any time period is simply a linearly homogeneous transformation of the same function for any other once the firm has implemented the technology implied by its rate/volume choice.

is arbitrary. The most important aspect of the rate-volume distinction is that it clarifies the relation of output to costs. It is not the time involved over which the rate of output is measured that determines costs but rather the input combination employed. A firm that may acquire factors of production in the appropriate amounts at prices that reflect their opportunity costs may produce at the same low cost for two months out of the year that another firm achieves with year-round production—even though the output per year of the latter is six times that of the former. In the next section we examine the implications of such intermittent production for industry structure.

IV. Coase & Economies of Scale

However far Alchian has taken us, his analysis still leaves a nagging question that has long plagued the profession. Robinson (pp. 19-20) says in discussing economies of scale,

> the small firm has a means of escape from the difficulty of attaining minimum efficient size, an escape very confusing to our attempt to analyse the structure of an industry. Where some given process requires a scale of production considerably greater than the smaller firms in an industry can achieve, this process tends to be separated off from the main industry, and all the smaller firms to get this particular process performed for them by an outside specialist firm.

Or put in Alchian's framework, if a firm has a planned volume of production that is insufficient to justify owning the most durable die, why can it not simply rent the durable die from someone else?

Cost with Discontinuous Production.

One way of formalizing this notion is to consider the firm breaking up its time horizon into periods of production and idle time.[13] The amount produced is adjusted by varying the length of the production run relative to the idle period. Figure 3 shows an average cost curve derived from Alchian's rate-volume map for the optimal planning period t_0. Also drawn is an average cost curve derived for the period $(2/3)t_0$. As we pointed out in the last section, these curves are simply proportional compressions and expansions of each other. The curve for $(2/3)t_0$ is just a reduced version in the horizontal dimension of that for t_0. The minimum average cost occurs at q_0 when production continues for t_0, and at $q_1=(2/3)q_0$ when the period of production is $(2/3)t_0$. Similarly, the average cost of producing q_1 units in t_0 time is AC_1, as is the average cost of producing $(2/3)q_1$ in $(2/3)t_0$ time.

It is obvious from the graph that the average cost of producing q_1 units in $(2/3)t_0$ is less than if production is spread over the entire period, t_0. Implicitly, stretching out production over t_0 involves the use of smaller less efficient machines, less labor specialization, etc. Producing q_1 units in $(2/3)t_0$ involves a technique of production that is as if q_0 was to be produced over the entire t_0 period. The scale effects that are enjoyed when the production target is q_0 are used to their full advantage for two-thirds of the time. If the instantaneous rate of production and the implied volume over the entire period t_0 is a free choice for the firm, the firm will clearly choose the technique and rate of production that minimizes average cost. In other words, the firm will employ specialized labor, durable

[13] See Maloney and McCormick.

dies, big machines and containers, and will plan production as if it were going to produce enough during t_0 to reach minimum average cost. Even though it only produces q_1, it acts like it is producing q_0. Economies of scale are fully exploited in this world and they pass out of the economic realm and exist only as textbook examples in engineering. The problem is, what happens in the idle time?

The Transaction Costs of Discontinuous Production

The firm employing this strategy of intermittent production does not do so at zero cost. There are storage costs and interest costs associated with holding inventories of finished goods, and the transaction costs of renting or idling capital vary with the time period of production.

Consider inventories. The model assumes that demand is continuous and uniform across t_0. Hence, the firm is forced to accumulate inventories in the period of production that are used to satisfy demand during the shut-down. Inventories are costly, so the firm may accept higher unit production costs to offset inventory costs. That is, the firm may not exploit intermittent production to the fullest extent, but may accept higher production costs to lower inventory costs.

Interestingly enough, there is an alternative strategy that mitigates inventory costs. Suppose that the firm has a contract for 8000 bobbins over the year, and that the average cost minimizing production schedule is 12,000 per year or 8000 in eight months. If the firm produces for eight months at 12,000 per year, it has accumulated an inventory of 2,667 at the end of the eighth month. However, by breaking production into two periods of six months, in which 4 are devoted to production and 2 to shut-down, inventories are reduced. The peak inventory level falls to 1333. Moreover, four periods of three months with two on and one off is even better. Peak inventories fall to 667 units. The sum of units held per day continues to fall as the number of intermittent production cycles increases in a given time period. In the limit of the fractal process, storage costs go to zero.[14]

However, this process of intermittent production may itself impose costs. Shut-downs and start-ups are not achieved instantly. Furthermore, the costs of retooling, reguaging, and cleaning may not be independent of the number of shut-downs per unit of output. Consider also the cost of renting as opposed to owning capital equipment. Here, again, multiplying and shortening the production runs increases costs. Transportation costs, for instance, would explode in the limit of moving a drill press back and forth across town for five minute sessions in different plants.

A strategy that mitigates all these problems and one that has a good deal of empirical importance is product diversification. Instead of the firm renting out its capital equipment (or renting someone else’s), it internally transfers them from one product to the next. The cost of inventories is, then, only weighed against the shut-down and start-up costs of switching product lines.[15] Thus, we expect to see firms diversifying into product lines that minimize switching costs. In the best example we have come across, Procter and Gamble produces three soaps, Tide, Cheer, and Oxidol. Oxidol's speckled crystals are the clean out change over from Cheer to Tide.

[14] The shape of the inventory pattern is repeated in smaller version as the production period is broken into smaller and smaller pieces. The process is described as a fractal; see Mandelbrot.

[15] The agency costs associated with renting are also eliminated.

We call the intermittent production view of economies of scale the *Coasian twist.* Ronald Coase is widely recognized as the grandfather of transactions costs.[16] He was one of the first economists to point out that in the neoclassical world absent transactions costs, firms are a redundant institution. Similarly, intermittent production implies that economies of scale only impact the behavior of firms because of the costs associated with moving resources in and out of production. Call these transactions costs and you have a Coasian twist on the whole problem.

Economies of scale are best explained in engineering terms through technical effects such as the economy of durable dies and the specialization of labor. However, it is the transaction costs associated with intermittent production that inform the empirical dimension of industry structure. Economies of scale can and are exploited in the rental market for many production processes. Construction companies that must move their production activity from place to place as a matter of course, routinely rent equipment. One of the transactions costs of the rental market is transportation and this cost is the same for construction companies whether they rent or buy. Hence, we should not expect economies of scale to be an important factor in the construction industry.

Similarly, firms within industries like machine tools and textiles will not exhibit economies of scale because of the ease of product diversification. A machine shop can make a variety of products from basic cutting, welding, bending, and extruding tools. Textile firms can switch the fiber characteristics across a virtually limitless array at very low cost. On the other hand, industries that use highly specialized resources that are also fixed in location are prime candidates for the sort of industry structure effects traditionally associated with economies of scale.

V. Summary

The purpose of this paper is to identify those characteristics that cause unit costs to fall as the size of the firm increases. Economies of scale come from a number of effects that the firm can take advantage of if the size of the production project suffices. Two important ones are the specialization or division of labor described by Smith and the economies of large machines and durable dies as discussed by Robinson. Workers become specialized in performing specific tasks if they are able to repeat them often enough. Specialization permits the work to be spread out among workers with fewer skills who command lower wages per task. Also, specialized labor is associated with assembly lines where tools and materials can be stretched out more conveniently. In all cases, the extent to which workers can be specialized is limited by the need to keep all workers fully employed. Large machines are more durable and typically require fewer resources in their construction and operation on a per unit of output basis. All of these principles are asserted to be states of nature without economic justification.

Importantly, these effects can be summarized in the dimensions of rate and volume as developed by Alchian. The rate effect increase average cost while the volume effect decreases it. Along the typical U-shaped average cost curve, the volume effect dominates early and the rate effect overwhelms later. The volume effect is the ability to plan to produce more in which case the firm chooses the more efficient, durable dies and

[16] "The Nature of the Firm" and "The Problem of Social Costs" are both founded on the notion of transactions costs. These are seminal works on the topic.

may employ more specialized labor. The rate effect is ultimately due to the depletion of specialized resources.

Finally, it is shown that economies of scale are a transactions cost phenomenon. If the firm chooses the most efficient machines and assembly processes, without the output requirements to justify them, the resources will be idled for some period of time. If the firm can diversify its product line or if the firm can rent (out or in) the durable machines, then the idle time problem is reduced, but it is replaced by the costs associated with renting resources and with diversifying the product line. These are essentially transactions cost problems which are the focus of extensive research spawned by the seminal research of Coase.

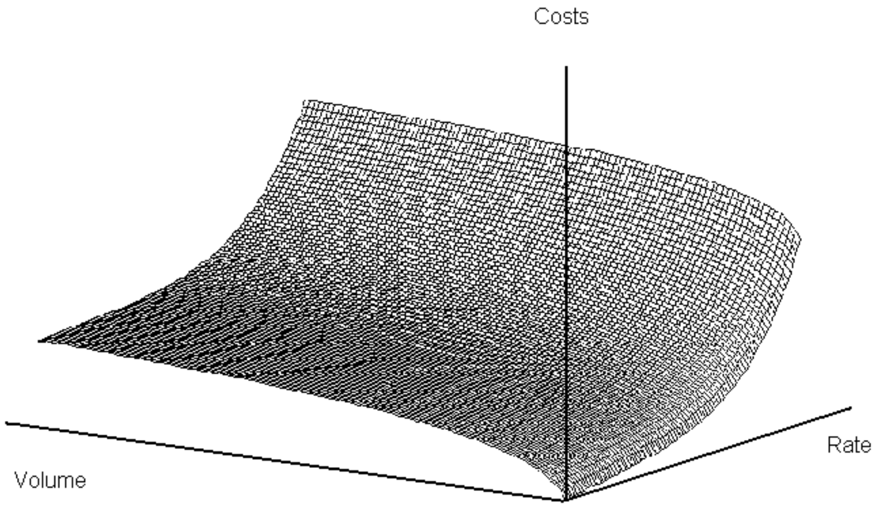

Figure 1: Alchian's Three Dimensional Cost Function

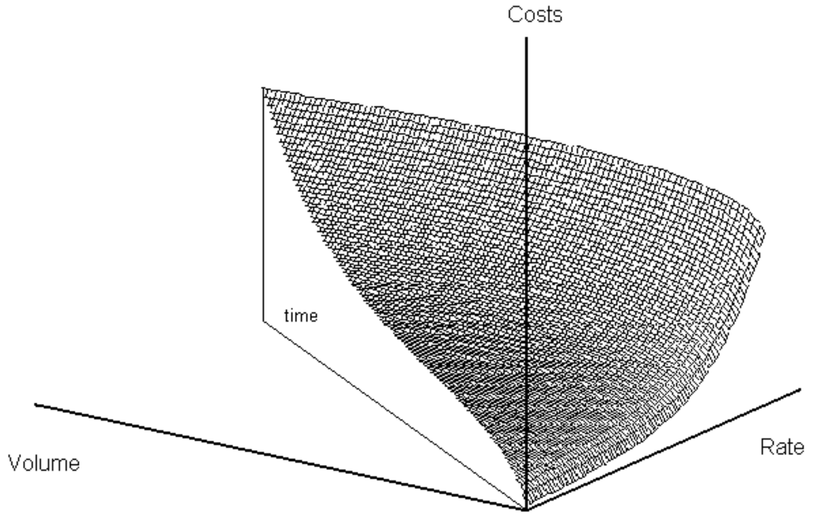

Figure 2: Alchian's Cost across a Constant Rate-Volume Ratio

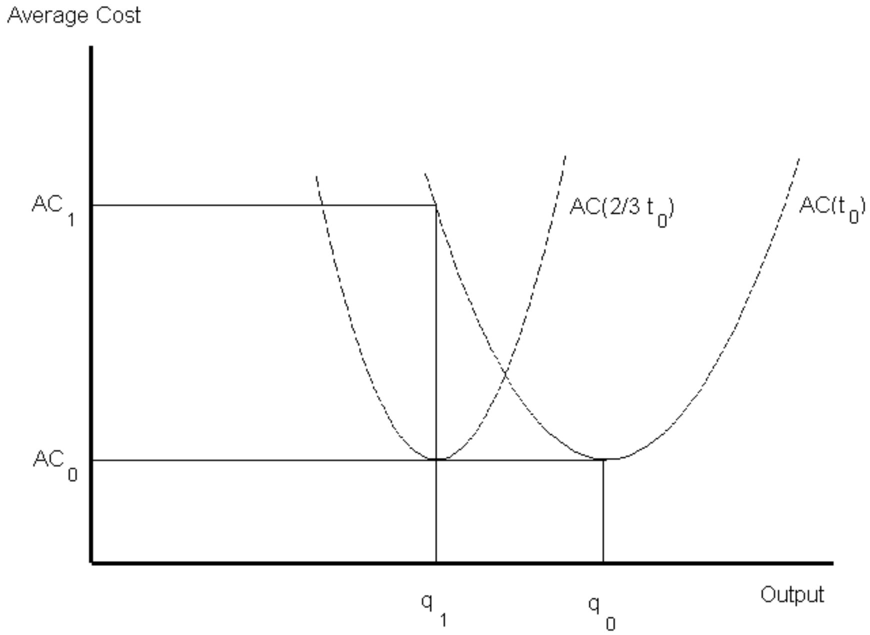

Figure 3: Average Cost with Intermittent Production

References

Armen Alchian, “Costs and Outputs,” in *Economic Forces at Work*, Indianapolis: Liberty Press, 1977.

Ronald H. Coase, “The Nature of the Firm,” *Economica*, 4 (1937), 386-405.

_______,“The Problem of Social Costs” *Journal of Law and Economics*, 3 (1960), 1-44.

MichaelT. Maloney and Robert E. McCormick, “A Theory of Costs and Intermittent Production,” *Journal of Business*, April 1983, 139-154.

Benoit B. Mandelbrot, *Fractals: Form, Chance, and Dimension*, San Francisco: W.H. Freeman, 1977.

E.A.G. Robinson, *The Structure of Competitive Industry*, Digswell Place: James Nisbet and Co. Ltd, 1931, Cambridge: University Press, 1958.

Adam Smith, *An Inquiry into the Nature and Causes of the Wealth of Nations*, New York: The Modern Library, 1937.

George J. Stigler, “The Division of Labor is Limited by the Extent of the Market,” *Journal of Political Economy*, June 1951.